Gene tree reconciliation: new developments in Bayesian concordance analysis with BUCKy

Cécile Ané[1,2], Colin N. Dewey[3,4], Satish Kumar Kotha[4], Bret R. Larget[1,2]
[1] Department of Statistics, University of Wisconsin-Madison
[2] Department of Botany, University of Wisconsin-Madison
[3] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison
[4] Department of Computer Sciences, University of Wisconsin-Madison

When different genes provide incongruent gene trees, some method for gene tree reconciliation is needed to extract the phylogenetic information from the sequence data. Bayesian Concordance Analysis aims to extract the vertical signal and to infer clusters of genes that share the same tree topology. Inference is made on the *genomic support* of clades, measured as the proportion of genes that truly have the clade in their tree, or the concordance factor of the clade. This genomic support is fundamentally different from the *statistical support* for a clade to be in the concordance tree, which is determined by the precision in estimated concordance factors. This is illustrated by the analysis of over 30,000 gene fragments in apes: BUCKy inferred a 1.0 posterior probability for the concordance tree (statistical support) even though one clade had a 76% concordance factor (genomic support), showing substantial true variability among gene trees. Bayesian concordance analysis, as implemented in BUCKy, does not make specific assumptions as to which biological processes are causing the potential gene tree discordance. Therefore, the estimated pattern of discordance can be used to test the null hypothesis that incomplete lineage sorting is solely responsible for the observed incongruence.

The concordance tree provides a summary of both the vertical phylogenetic signal (the tree topology) and the presence of horizontal signal (from the concordance factors of clades). A fast way to reconstruct the concordance tree is to build it from the clades with highest concordance factors. However, it was shown that this fast method could return an inconsistent estimation of the vertical signal in some cases when the discordance between gene trees is caused by the coalescent process only. A consistent alternative is to reconstruct the population tree based on all the quartets with highest concordance factors. The recent developments made to the program BUCKy include this quartet-based reconstruction of the population tree. Simulations show the consistency of this estimated population tree when gene trees are generated according to the coalescent process along species trees with very short internal branch lengths.

Program distributed under the GNU general public license at
www.stat.wisc.edu/~ane/bucky/

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution 24:412–426.
Baum, D. A. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon 56:417–426.
Ané, C. 2010 (in press). Reconstructing concordance trees and testing the coalescent model from genome-wide data sets. In Estimating Species Trees: Practical and Theoretical Aspects (Knowles L. L. and Kubatko L., eds.) Wiley-Blackwell, Hoboken, NJ.